The assessment of clinical competence in high stakes examinations: are we justified in abandoning old methods in favour of the new ?

Summary

Chapter 1: Background

There is increasing pressure for professional accountability and assurance to the public that doctors are clinically competent to practise. Medical schools in the UK, in order to deliver more reliable examinations which test broadly across the curriculum, have moved away from traditional assessments such as long cases and orals to more "atomised" formats such as the Objective Structured Clinical Examinations (OSCEs) and Multiple Choice Questions. Yet this change has occurred in the face of a paucity of published psychometric data on traditional methods. This thesis studies a final examination in the process of change from old methodology to new and seizes the opportunity to investigate, the psychometrics of the traditional formats in more depth, before their demise. The key research question overarching the thesis is "Have we become too "atomistic" in our approach to assessing clinical competence? Are we justified in abandoning the more integrated traditional high stakes examinations formats, in particular the long case, too early without either attempting to improve their validity or measuring their reliability?"

Chapter 2

A review of the important principles underpinning the assessment of clinical competence, which aimed to achieve an international perspective, introduces the thesis. The importance of carefully blueprinting the content of examinations against the curricula, setting appropriate standards, balancing the validity of the tests against their reliability and weighing the formative and summative functions of the assessment are outlined. The paper concludes that current examination formats tend to focus too heavily on knowledge based competencies and further research into the format and reliability of composite medical examinations is important.

Chapter 3

A battery of tests is often used to measure all the attributes of clinical competence. The final medical school examination under study focused on assessing competency traits; "knowledge, skills and attitudes". It consisted of three written papers, an OSCE and two history taking long cases. The initial research question asked "How reliable is a composite final year undergraduate examination?" Using multivariate generalisability theory, variances in test length and composition were accounted for and an index of overall reliability was obtained. The results highlighted that careful structuring of papers to balance the length and format of individual tests is crucial to ensure acceptable and meaningful reliability is achieved.

Chapter 4

Previous studies had looked at the inter-rater reliability of long cases but there was little data on inter-case reliability. By including two modified history taking (HT)

long cases with real patients alongside the OSCE, generalisability theory was used to answer the question "How many long cases would be needed to achieve a reliable high stakes test?" It was concluded that eight to ten thirty minute HT long cases would achieve reliability consistent with a high stakes examination. Thus provided sufficient testing time and real patients were available to cover a range of content specificity, the long case could not be abandoned for reasons of reliability alone. Length of testing time rather than the method used appears to reign supreme.

Chapter 5

In the traditional long case format only the presentation of the patient is marked and the interaction between patients and student is not observed. A further modification of the examination was set in place to test the hypothesis that if the patient-student interaction was observed and graded a more valid assessment would result. Ratings of two examiners observing and marking the student taking the history were compared with the independent marks of two different examiners assessing the presentation. Ratings of observation and presentation contributed significantly and independently to the correlation with clinical competence, as judged from the candidates' OSCE scores. Observing the interaction proved worthwhile.

Chapter 6

A further research question arose. If extending testing time improved the reliability of long cases, would the same principle apply to traditional oral examinations? Generalisability theory was used to analyse the oral components of the Membership of the Royal College of General Practitioners examination. Although some attempt at improving the content of the orals had been made, examiners continued to use their own questions in the traditional way. The same principle held true. Extending testing time to four 20-minute orals, each with two examiners, or five orals, each with one examiner, resulted in predicted pass/fail reliabilities consistent with high stakes tests.

Chapter 7

Finally when weighing the advantages of the new approach of using simulations rather than real patients, we developed concerns that the use of simulations was not achieving true objectivity within the OSCE. We were concerned that, as ethnic minority students tended to perform less well in the OSCE, there was possible disadvantage either in the setting of the stations or the marking. Despite standardisation and itemisation of communication skill stations in an OSCE was there still potential for bias in ratings? Interactions on four communication stations were videoed and analysed using qualitative discourse analysis. No examiner bias was seen but the research highlighted the importance of ensuring stations are carefully constructed so that ethnic minority candidates are not disadvantaged.

Conclusions

These studies have shown that it is not the test format that matters as much as the test content i.e., as long as the examination samples widely enough, traditional formats have the same potential as new ones to provide reliable assessments of clinical competence. The trait model of approaching assessment expressed as domains of knowledge, skills and attitudes may be too simplistic and can create complicated tests which need careful design if the purpose of the test is to be achieved and bias avoided. Indeed we need to be cautious when thinking in terms of behaviourist traits as these tend to atomise the process.

No single method is flawless and the advantages of using standardised rather than real patients may not be as significant as was originally thought. The traditional assessment methods, such as the long case and orals, use a more integrated process and should not be discarded although the feasibility of using these within summative assessments is such that they have a more formative rather than summative role. The current move away from examinations towards more assessment in the workplace presents welcome opportunities to revisit and develop these test methodologies.

We need more research into designing and validating longitudinal assessments which use these integrated methods in a formative purposive way and which emphasise to students the educational importance of experience in interaction with patients.

Samenvatting

Hoofdstuk 1: Achtergrond

Er wordt steeds meer druk uitgeoefend op artsen om verantwoording af te leggen voor hun klinisch handelen. Ook worden steeds vaker garanties gevraagd dat artsen competent zijn om hun beroep uit te oefenen. Gezien deze ontwikkeling, hebben medische faculteiten gezocht naar methoden om klinische competentie betrouwbaar en gedurende het gehele curriculum te kunnen meten. Dit heeft ertoe geleid dat traditionele toetsmethoden, zoals 'long cases' en mondelinge examens, overboord zijn gezet en vervangen door 'geatomiseerde' toetsmethoden, zoals stationsexamens en meerkeuzetoetsen. Ten tijde van deze verandering waren er slechts weinig gepubliceerde gegevens over psychometrische eigenschappen van traditionele toetsmethoden beschikbaar.
Het in dit proefschrift beschreven onderzoek is uitgevoerd tijdens het overgangsproces van oude naar nieuwe toetsmethoden bij een bestaand afsluitend examen aan het eind van de basisartsopleiding. Dit overgangsproces is aangegrepen voor een diepgaand onderzoek van de psychometrische eigenschappen van de traditionele toetsmethoden voordat ze definitief werden afgeschaft. De centrale onderzoeksvraag is: Is onze benadering van toetsing van klinische competentie te atomistisch geworden? Is het verdedigbaar om traditionele geïntegreerde toetsmethoden, zoals de 'long case', voor belangrijke examens voortijdig af te schaffen, zonder te onderzoeken of de validiteit verbeterd kan worden en de betrouwbaarheid vast te stellen?

Hoofdstuk 2

Als inleiding wordt, vanuit een internationale invalshoek, een overzicht gegeven van de belangrijkste basisprincipes van toetsing van klinische competenties. Er wordt ingegaan op het belang van blauwdrukken om toetsing te laten aansluiten bij de curriculuminhoud, het vaststellen van beoordelingsnormen, het afstemmen van validiteit en betrouwbaarheid van toetsen en afwegingen ten aanzien van formatieve en summatieve toetsing. De conclusie is dat de huidige toetsmethoden te veel gericht zijn op competenties op basis van kennis en dat er onderzoek gedaan moet worden naar methoden en betrouwbaarheid van samengestelde examens in de geneeskundeopleiding.

Hoofdstuk 3

Vaak wordt een testbatterij gebruikt om alle aspecten van klinische competentie te meten. Het in dit hoofdstuk beschreven onderzoek betreft een afsluitend examen dat voornamelijk gericht is op toetsing van competentieaspecten: "kennis, vaardigheden en attitudes". Het examen bestaat naast drie schriftelijke onderdelen uit een stationsexamen en twee 'long cases', waarbij studenten de anamnese afnemen bij een echte patiënt. De onderzoeksvraag is: "Hoe betrouwbaar is een afsluitend samengesteld examen in het laatste jaar van de basisartsopleiding?". Met behulp

van multivariate generaliseerbaarheidstheorie zijn de effecten van variatie in toetsduur en toetssamenstelling onderzocht en is een globale betrouwbaarheidsindex opgesteld. Uit de resultaten blijkt dat afstemming van duur en vorm van afzonderlijke toetsonderdelen een cruciale voorwaarde is voor acceptabele, zinvolle en betrouwbare toetsing.

Hoofdstuk 4

Aan de interbeoordelaarsbetrouwbaarheid van 'long cases' zijn verschillende onderzoeken gewijd, maar over intercasusbetrouwbaarheid is maar weinig bekend. Twee aangepaste 'long cases', waarbij studenten de anamnese afnemen bij echte patiënten zijn toegevoegd aan een stationsexamen. Vervolgens is met behulp van generaliseerbaarheidstheorie onderzocht hoeveel 'long cases' nodig zijn voor betrouwbare toetsing. De resultaten laten zien dat bij een examenduur van 8 tot 30 minuten de betrouwbaarheid acceptabel is voor een afsluitend examen. Het blijkt dus dat betrouwbaarheid onvoldoende aanleiding vormt om het traditionele examen af te schaffen, mits er genoeg toetstijd en echte patiënten beschikbaar zijn voor voldoende inhoudelijke variatie in de toets. Niet toetsmethode maar toetsduur blijkt van doorslaggevend belang.

Hoofdstuk 5

De beoordeling van de traditionele 'long case' is uitsluitend gebaseerd op de patiëntenpresentatie door de student, terwijl het student-patiëntcontact niet geobserveerd wordt. Een verdere aanpassing van het examen maakte het mogelijk om de hypothese te toetsen dat de validiteit van de beoordeling verbetert als het student-patiëntcontact geobserveerd en beoordeeld wordt. Het oordeel van twee examinatoren die de studenten tijdens de anamnese observeren en beoordelen, is vergeleken met het oordeel van twee verschillende examinatoren op basis van patiëntenpresentaties. Beide oordelen blijken een significante en onafhankelijke bijdrage te leveren aan de correlatie met de beoordeling van klinische competentie op basis van het stationsexamen. Het blijkt dus zinvol om patiëntencontacten te observeren.

Hoofdstuk 6

Een andere onderzoeksvraag diende zich nu aan. Als een langere toetsduur de betrouwbaarheid van 'long cases' vergroot, doet eenzelfde effect zich dan ook voor bij traditionele mondelinge examenvormen? Met behulp van generaliseerbaarheidstheorie zijn de mondelinge onderdelen van de toelatingstoets voor het Royal College of General Practitioners geanalyseerd. Ondanks pogingen om de inhoud van de mondelinge examens te verbeteren, hielden de examinatoren vast aan de gebruikelijke manier van examineren. De analyse laat inderdaad een vergelijkbaar effect zien als bij de 'long case'. Verlenging van het examen tot vier mondelinge

toetsen van twintig minuten met twee examinatoren of tot vijf mondelinge toetsen met vijf verschillende examinatoren leidt tot een voorspelde betrouwbaarheid die acceptabel is voor een belangrijk examen.

Hoofdstuk 7

Naar aanleiding van afweging van de voordelen bij de nieuwe toetsaanpak van simulaties en echte patiënten, rees de vraag of simulaties in het stationsexamen wel voldoende garantie bieden voor echte objectiviteit. Aanleiding tot bezorgdheid waren de minder goede prestaties van allochtone studenten. Deze vormden een aanwijzing dat er factoren in de setting van de stations of de beoordeling waren, die in het nadeel van deze studenten konden werken. Wordt de beoordeling van stations over communicatievaardigheden, ondanks standaardisatie en gedetailleerde beoordelingslijsten, negatief beïnvloed door vooroordeel? De interacties bij vier stations over communicatievaardigheden zijn op video opgenomen en onderzocht met behulp van kwalitatieve tekstanalyse. Er werd geen vooroordeel bij de examinatoren waargenomen, maar de onderzoeksbevindingen onderstrepen het belang van uiterste zorgvuldigheid bij het ontwerpen van stations om effecten te voorkomen die nadelig kunnen zijn voor kandidaten van allochtone afkomst.

Conclusies

Dit proefschrift toont aan dat de toetsmethode van minder doorslaggevend belang is dan de toetsinhoud. Met andere woorden: bij een voldoende gevarieerde exameninhoud kan klinische competentie even betrouwbaar gemeten worden met traditionele als met nieuwe toetsvormen. Een model dat onderscheid maakt tussen toetsing van kennis, vaardigheden en attitude is wellicht te eenvoudig. Dit kan leiden tot ingewikkeld toetsen die zeer zorgvuldig samengesteld moeten worden om het beoogde toetsdoel te kunnen bereiken en vooroordeel uit te sluiten. Grote voorzichtigheid is geboden bij het denken in termen van gedragsmatige aspecten, omdat dit kan leiden tot geatomiseerde toetsprocedures.
Geen enkele toetsmethode is volmaakt en wellicht bieden gestandaardiseerde patiënten in vergelijking met echte patiënten minder grote voordelen dan aanvankelijk werd gedacht. Traditionele toetsmethoden, zoals 'long cases' en mondelinge examens, worden gekenmerkt door een geïntegreerde benadering. Afschaffing van deze examens lijkt ongewenst, ook als in aanmerking wordt genomen dat zij uit haalbaarheidsoverwegingen geschikter zijn voor formatieve dan voor summatieve beoordeling. De huidige tendens om traditionele toetsvormen te vervangen door beoordelingen in de praktijk biedt een uitstekende aanleiding om deze toetsmethoden te herontdekken en verder te ontwikkelen.
Onderzoek is nodig naar de ontwikkeling en validering van longitudinale geïntegreerde beoordelingsmethoden voor formatieve doeleinden, die bovendien de leerzaamheid van contacten met echte patiënten kunnen benadrukken.